caBIG cancer Biomedical Informatics Grid

DRAFT DRAFT DRAFT DRAFT

April 5, 2005 DRAFT

caBIG[TM] Compatibility Guidelines Revision 2

The Cancer Biomedical Informatics Grid[TM] Program

TABLE OF CONTENTS

INTRODUCTION

Purpose

The purpose of this document is to provide the cancer Biomedical Informatics Grid^{TM} (caBIG^{TM}) community with compatibility guidelines for creating software systems that are syntactically and semantically interoperable. The guidance contained herein is intended to support the evaluation of existing systems and to inform the designs of new systems. This document focuses on issues related to the representation of, access to, and exchange between biomedical informatics resources. Requirements for integration and use of the caBIG standards management infrastructure are also addressed. However, with few exceptions, a particular technology implementation of a given system or tool is not specified.

caBIG

caBIG is a voluntary network or 'grid' of individuals and institutions that are working to create a better environment for the sharing of cancer research data and software tools. The goal is to speed the delivery of innovative approaches for the prevention, detection and treatment of cancer. The infrastructure and tools created by caBIG also have broad utility outside the cancer community. caBIG is being developed under the leadership of the National Cancer Institute and its Center for Bioinformatics.

caBIG Principles and Implications for Interoperability

The caBIG program has defined several principles that have implications for interoperability and for the creation and dissemination of the compatibility guidelines themselves:

- Open Source/Open Access. Products that are funded by NCI in connection with the caBIG initiative must be made available under licenses that permit unrestricted use and redistribution by any party, whether commercial, academic, or non-profit. Therefore, these compatibility guidelines and any resources or specifications related to caBIG interoperability standards must also be distributed according to these terms.

  Note, however, that groups can develop systems and tools that implement caBIG standards, and thus meet all necessary compatibility requirements, without necessarily providing the resulting products under an open source/open access license, as long as this development was not funded by the caBIG program.

- Open Development. caBIG funded activities must be conducted in open forums, with opportunity for observation, comment, and contribution by any interested member of the community. Therefore, the process by which these guidelines and any related interoperability standards are created must provide for public involvement, comment and review.

- Federated. The caBIG program envisions a federation of cancer biomedical informatics resources, not a single repository or hosting center. Therefore, these compatibility guidelines must be sufficiently rigorous to enable developers of independently managed information resources and tools to achieve system interoperability with systems not under their direct control.

Levels of Maturity

The caBIG community has recognized there can be differing degrees of interoperability between systems, and that these can be qualified in terms of maturity level. The caBIG Compatibility Guidelines are thus organized into four levels of maturity: Legacy, Bronze, Silver, and Gold.

- Legacy. Implies no interoperability with an external system or resource. A system that was designed either without awareness or prior to the availability of these compatibility guidelines, and which does not meet any of the requirements for interoperability.
- Bronze. Classifies the minimum requirements that must be met to achieve a basic degree of interoperability.
- Silver. A rigorous set of requirements that, when met, significantly reduce the barrier to use of a resource by a remote party who was not involved in the development of that resource.
- Gold. Currently being defined by the caBIG participants. Is expected to provide for a formalized grid architecture and data standards that will enable standardized advertising, discovery, and use of all federated caBIG resources.

Interoperability Definitions and Goals

Interoperability can be defined as the ability of a system to access and use the parts of another system. The caBIG program has made interoperability between data and software components a primary strategic goal. These compatibility guidelines provide a high-level description of the decisions made to date with respect to requirements for interoperability. The cross-cutting Architecture and Vocabulary/Common Data Elements Workspaces were created as part of the caBIG initiative to provide an ongoing forum and mechanism for defining and ensuring interoperability across caBIG technology and data products. The activities of these workspaces will result in more detailed standards specifications and requirements, thus ensuring that the program goals are met.

It is useful to consider the interoperability requirements for access independently from those for usage, though of course they must be synthesized in the final implementation. “Access” requirements in caBIG include programmatic access to data and tools from software, not just interactive access from end-user interfaces. Given this requirement, the primary obstacle to “accessing” parts of another system is the heterogeneity in the programming and messaging interface syntax across systems that have been developed by independent groups, if indeed these interfaces exist at all. The problem of access is therefore a problem of poor syntactic interoperability. Regularization of application programming and messaging interfaces is necessary to overcome obstacles to syntactic interoperability.

“Use” of a resource demands more than just access. Scientific analysis and interpretation requires a deep understanding of the procedures, manipulations and parameters that go into the creation of a data resource or tool. Given this requirement, the primary obstacle to “using” parts of another system is the ambiguity behind the origins and meaning of the data. The problem of usage is therefore a problem of poor or ambiguous semantic interoperability. Explicit descriptions and definitions of the contents and meanings of resources are necessary to overcome barriers to semantic interoperability.

The highest degree of interoperability is attained when access and use can be completely automated. To achieve this level of interoperability, programming and messaging interfaces must conform to standards that specify consistent syntax and format across all systems in the federation. Further, all data must be associated with metadata and terminology identifiers and codes that support computational aggregation and comparison of information that resides in separate resources.

Achieving Syntactic and Semantic Interoperability

When considering how to overcome the obstacles to interoperability, the caBIG program members arrived at four areas that need to be addressed. One of the four address issues related to syntactic interoperability; the remaining three address issues related to semantic interoperability. The four areas are:

- Programming and Messaging Interfaces. Computer programs and the people who write them are able to access resources from other programs through programming and messaging interfaces. Each of these interfaces responds to a particular syntax for its communications. Agreement upon standards for these interfaces is necessary to overcome barriers to syntactic interoperability.
- Vocabularies and Ontologies. Biomedical information includes a substantial body of specialized concepts that are represented by terms. Agreement upon the basic concepts, terms and definitions that are inherent in all biomedical information is essential for achieving semantic interoperability. Terminology development systems that use description logic are helpful tools for managing these concepts.
- Common Data Elements. Data that is collected on a given study or trial must be defined and described such that remote users of that data can understand what it means. These metadata descriptions are referred to as data elements. When many groups use the same (common) data elements (CDEs), then larger-scale studies can be conceived, since consistency and comparability of across sites, studies, and time becomes possible. CDEs are therefore critical constructs for semantic interoperability.
- Information Models. Individual types of data are rarely collected or presented in isolation. Rather, they are assembled into a contextual environment that includes closely and more distantly associated data and information. These associations and relationships can be presented in the form of an information model. These models convey both a human and a machine understandable representation of the contextual environment of data in an information resource, and are important for achieving the highest degree semantic interoperability.

COMPATIBILITY MATRIX

This Compatibility Matrix is a summary of caBIG compatibility requirements. Please refer to the body text of the document for complete information on compatibility.

Maturity Model	Legacy	Bronze	Silver	Gold
Programming and Messaging Interfaces	*- No programmatic interfaces to the system are available. Only local data files in a custom format can be read* *- Data transfer mechanisms implemented only on an ad hoc basis*	*- Programmatic access to data from an external resource is possible.*	*- Well-described API's provide access to data in the form of data objects.* *- Standards-based electronic data formats are supported for both input to and output from the system.* *- Standards-based messaging protocols are supported wherever messaging is relevant.*	*- All features of Silver, plus:* *- Service-oriented components produce or use resources in the form of grid services* *- Interoperable with data grid architecture to be defined by caBIG*
Vocabularies / Terminologies & Ontologies	*- Free text used throughout for data collection*	*- Use of publicly accessible controlled vocabularies as well as local terminologies.* *- All terminologies must include unambiguous definitions of terms*	*- Terminologies reviewed and validated by the caBIG Vocabulary/Common Data Element (VCDE) Workspace used for all appropriate data collection fields.*	*- All features of Silver, plus:* *- Full adoption of caBIG terminology standards as approved by the VCDE workspace*
Data Elements	*- No Structured metadata is recorded*	*- Data element descriptions are maintained with sufficient definitional depth to enable a subject matter expert to unambiguously interpret the contents of the resource without contacting the original investigator.* *- Data elements are built using controlled terminology* *- Metadata is stored and publicized in an electronic format that is separate from the resource that is being described..*	*- Common Data Elements (CDEs) built from controlled terminologies and according to practices validated by the VCDE workspace are used throughout.* *- CDEs are registered as ISO/IEC 11179 metadata components in the cancer Data Standards Repository (caDSR)*	*- All features of Silver, plus:* *- CDEs designated as caBIG Standards by the VCDE workspace are used* *- Metadata is advertised and discoverable via the caBIG grid services registry*
Information Models	*- No model describing the system is available in electronic format*	*- Diagrammatic representation of the information model is available in electronic format.*	*- Information models are defined in UML as class diagrams and are reviewed and validated by the VCDE Workspace.*	*- All features of Silver, plus:* *- Information models are harmonized across the caBIG Domain Workspaces*

Programming and Messaging Interfaces

The compatibility criterion of ‘Programming and Messaging Interfaces' addresses issues related to programmatic access to a resource, input and output formats, and messaging protocols. To achieve Bronze compatibility, the resource should provide baseline programmatic access to data. That is, a public, documented Application Programming Interface (API). The API needs to be rich enough to provide basic query and retrieval of information. This requirement does not place specification on the specific technology used to create and propagate the API.

Silver-level compatibility is far more demanding. Data-oriented systems must provide a well-documented public API that is based upon an object-oriented abstraction of the underlying data. This abstraction layer must be derived from an information model constructed in the Unified Modeling Language (UML; see *Information Models* below). Data must be returned in the form of data objects that are instances of classes in the model. Data formats must conform to standards set by the caBIG workspace with which the resource is aligned. Wherever relevant, standards-based messaging systems are used to exchange information. Silver-level analytical tools and client applications must be able to read directly from these caBIG-compatible APIs.

Gold-level Programming and Messaging Interfaces are currently being defined by the caBIG program participants. Several decisions have been made to date: The Gold architecture will include a service-oriented data and analytical service grid with standardized service advertising and discovery features; service APIs will communicate using a specified XML syntax, and will return results as data objects that have been serialized into XML; an identifier system for data objects will be implemented across all grid data services; a grid-level security strategy will be implemented to allow for access control.

Vocabularies/Terminologies and Ontologies

An important feature of modern terminology management is the recognition that the "concept" is the unit of semantic meaning, not simply the term or word. Concepts are described by preferred terms, synonyms, definitions and other properties. Given the diversity and overlap in meaning of terms in use, it is useful to use description logic to create and maintain concepts and to describe the relationships among concepts. These frameworks support the production of thesauri of non-redundant concepts that can be used to implement terminological and semantic consistency in data systems.

At the Bronze level of maturity, the information resource utilizes public controlled vocabularies in parts of the data collection and reporting process, but may supplement them with local vocabularies. All terminologies, including those developed locally, must include unambiguous definitions of terms.

Silver-level maturity introduces the requirement for review and approval of terminologies by the caBIG VCDE workspace. Local or private terminologies that are not available to the caBIG community may not be implemented. Gold compatibility is similar to Silver, but with the added requirement that registered standards approved for caBIG-wide usage are implemented wherever they are available.

Given the dynamic nature of scientific research, terminology standards for caBIG are expected to grow and evolve as the scope of the program grows. Therefore, the enhancement and extension of currently available terminology sources is anticipated to be an ongoing activity.

Data Elements

While controlled terminology sources provide the semantic "raw material" for interoperability, they are stand-alone, independent resources that do not describe any particular data system. Developers of data management systems must separately characterize the contents of the actual system by mapping the data fields to structured metadata, or data elements. This requirement for documenting the metadata only covers attributes exposed as part of the system's public APIs, not all the internal features of lower layers.

Common Data Elements provide a means toward semantic continuity and data comparability across studies over time. CDEs help solve problems of ambiguity by providing precise definitions of data fields and types, sufficient to unambiguously characterize the specific meaning of any particular datum collected in a research study. CDEs ultimately save analysis time by minimizing the need to reverse engineer meaning from data, and also enable consistent data collection across locations in large multi-site investigations.

Bronze-level systems have their metadata structured into an electronic format that details the specification of each data element that is in the system. These metadata are constructed from the selected controlled terminology sources, and include sufficient descriptive information to enable a subject matter expert to interpret the contents of the system without having to contact the original investigator. The metadata are exposed in a publicly accessible electronic resource that is distinct from the information system itself.

Silver is once again more rigorous, but as such provides for a much higher degree of semantic interoperability, including the provision for computational aggregation and comparison of data. Common Data Elements constructed according to best practices defined by the caBIG VCDE workspace must be used. These CDEs are all registered in the caBIG Context of the NCI cancer Data Standards Repository (caDSR), an implementation of the ISO/IEC11179 standard for metadata registries. Reuse of existing validated CDEs in the caDSR must be considered before any new data elements are created. All new CDEs are subject to review and validation by the VCDE workspace before they are deployed.

It is worth noting that there are two major mechanisms for creating CDEs in the caDSR. Editing tools that operate directly on the caDSR can be used by trained metadata curators to construct individual data elements and their associated components. The other alternative is to derive data elements from a properly constructed information model in UML. Such models can be submitted by caBIG projects for loading by caDSR database administrators.

Gold requirements for metadata will likely be an extension of the Silver specification, but with the added requirement that CDE standards approved for caBIG-wide usage are implemented. There will likely be with additional provisions added for metadata advertising, data provenance, and identifiers in the caBIG grid architecture.

Information Models

Data Elements are precise specifications of individual types of data that are collected during a research study. However, scientific interpretation relies on the placement of data elements into a broader semantic context, an information model. Therefore, in order to attain the highest degree of semantic interoperability, data must be expressed in the context of such a model.

The Bronze-level requirement for an information model is quite modest. A diagrammatic representation of the information structure that is being produced by a system is necessary, and must be available in an electronic format.

Silver-level compatibility requires the use of the industry-standard modeling language, UML. In particular, UML class diagrams that illustrate the data classes, attributes, and relationships is required. (Using other aspects of UML modeling is encouraged as a best practice in development methodology, but is not central to the issue of semantic interoperability). Class diagrams must conform to the naming conventions and terminology standards prescribed by the caBIG program. UML models must be fully annotated with class and attribute definitions, and with associated terminology concept codes. Upon review and validation by the VCDE workspace, models can be submitted for registration and loading into the caDSR.

The benefits of using a standard modeling language are significant. UML is derived from a structured meta-model, and therefore all UML models share a common parental meta-structure. This trait allows for programmatic access to the models themselves, a feature that is leveraged when models are loaded into the caDSR. The common meta-model also enables software code to be automatically generated from the models, a key benefit of the model-driven architectural paradigm espoused by the Object Management Group. In this way, caBIG Silver requirements for Programming Interfaces can be satisfied by automatically generating model-driven middleware and database code.

Gold requirements for Information Models will likely involve an added degree of harmonization across caBIG domains.

CHANGES SINCE THE LAST REVISION

- Revision 1 of this document included a number of example system architecture diagrams that were intended to illustrate possible ways to deploy caBIG-compatible systems. These diagrams proved confusing, and distracted from the main theme of syntactic and semantic interoperability, and thus have been removed.
- "Interface Integration" has been renamed "Programming and Messaging Interfaces" to improve the clarity and precision of this label.
- Use of Common Data Elements registered in the caDSR is now required for Silver-level compatibility.
- Data Elements with sufficient definitional information to enable a subject matter expert to unambiguously interpret the the contents of the resource are now required for Bronze-level compatibility.
- Explanatory information has been revised and reorganized according to the four areas of compatibility rather than by Bronze-Silver-Gold classification.
- Initial features of the anticipated grid service-oriented Gold-level architecture are described.

USEFUL LINKS AND RESOURCES

- caBIG Architecture Workspace: http://cabig.nci.nih.gov/workspaces/Architecture. Forum for discussing, prototyping and defining caBIG architectural standards, interoperability technologies, and engineering best practices.
- caBIG VCDE Workspace: http://cabig.nci.nih.gov/workspaces/VCDE. Forum for establishing and reviewing the use of caBIG data standards.
- Introduction to Unified Modeling Language: http://www.omg.org/gettingstarted/what_is_uml.htm.
- NCI Center for Bioinformatics Core Infrastructure: http://ncicb.nci.nih.gov/core. Home of caCORE, NCI's information technologies and services for semantics and data management.
- Cancer Data Standards Repository: http://ncicb.nci.nih.gov/core/caDSR. caCORE component that hosts common data elements.
- Common Data Element Browser: http://cdebrowser.nci.nih.gov. Web application that provides CDE search, browse and retrieval capabilities.
- NCI Enterprise Vocabulary Services: http://ncicb.nci.nih.gov/core/EVS. caCORE component that provides terminology management and development services to the cancer community. Jointly managed by the NCI Center for Bioinformatics and Office of Communications.
- NCI Terminology Browser: http://nciterms.nci.nih.gov. Web application the provides browse and search capabilities for NCI Thesaurus and other terminologies.
- NCI Metathesaurus Browser: http://ncimeta.nci.nih.gov. Web application that provides browse and search capabilities for NCI Metathesaurus.
- caCORE Software Development Kit: http://ncicb.nci.nih.gov/core/SDK. Developer tools that assist with the creation of a caCORE-like system that meets caBIG Silver-level compatibility guidelines.